
(https://colab.research.google.com/github/geesanchez/desktopAI/blob/main/r_exercise_3.ipynb

# R Exercise 3

In [ …

```
# Instrumental Variables example
install.packages("sem")
library(sem)
wages<-read.csv('http://inta.gatech.s3.amazonaws.com/wage2.csv')
iv.results<-tsls(lwage ~ educ + age + married + black, ~ feduc + age + marriec
# Run an instrumental variables regression. Note that father's education is
# used for an instrument, and is not in the first set of variables. Age,
# marital, status, and race are assumed to be less related to underlying
# ability and so are controlled for in the "first stage" regression too.

ols.results<-lm(lwage ~ educ + age + married + black, data=wages)
# Run the analogous OLS regression without father's education

print(summary(ols.results))
print(summary(iv.results))
# Note that the point estimate for education has now gone up. This suggests
# that people who have higher education in a way that's correlated with
# their father having higher education, even holding fixed age, marital
# status, and race, receive ~9% higher wages for every year of education
# Also note that the standard error on education is higher for the iv than
# ols results. The reason for this is that in IV, we're using only some of
# the variation in education: only the part related to father's education,
# so it's like there's less variation overall

# Stepwise regression and tree example
install.packages(c('MASS','sem'))
library(MASS)
start.model<-lm(wage ~ hours + IQ + KWW + educ + exper + tenure + age + marrie
# Give an initial model, which will be the most coefficients we'd want to ever
summary(start.model)
stepwise.model<- step(start.model)
# The command "step" adds and subtracts coefficients to maximize a measure of
# goodness of fit, by default AIC
summary(stepwise.model)

# install.packages('tree')
library(tree)
wage.tree <- tree(wage ~ married + hours + IQ + KWW + educ + tenure + exper, m
# fit a tree
summary(wage.tree)
# Print the results

plot(wage.tree)
text(wage.tree)
# Plot the tree, and add the text decisions

cross.validation <- cv.tree(wage.tree)
# perform cross-validation, 10-fold.

cross.validation
# look for the size with the lowest "dev" or deviance
# Why might this be different than the fitted tree?

pruned.wage.tree<-prune.tree(wage.tree,best=5)
# Prune this tree down to 5 terminal nodes, just to show this is how you would
# do it. Here this makes it worse though
summary(pruned.wage.tree)

lmfit<- lm(wage ~ married + hours + IQ + KWW + educ + tenure + exper, data=wag
# Compute a regression using those same variables:
anova(lmfit)
summary(wage.tree)
# Notice that the mean sum of squared residuals was 130895 using the linear
# model, compared to the residual mean deviance of 130000 using trees. So the
# tree method has less residual error, which means it's a better predictor.
# This is especially striking given that the tree only uses KWW, educ, IQ,
# While the regression needs substantial contributions from tenure, experience
# and marital status too.

predict(wage.tree, newdata=data.frame(IQ=100, KWW=50, educ=16, married=1, hour
predict(lmfit, newdata=data.frame(IQ=100, KWW=50, educ=16, married=1, hours=40
# These predictors give different estimates for the expected wage of a
```

```
# particular individual, relatively young, married, well educated, with good
# knowledge of the world of work. Which would you trust and why?

# Extension: try to build another tree, and see what it looks like.
# You can sample the data with the command:
# install.packages('dplyr')
# library('dplyr')
# wages.sample <- sample(wages, n=500)

# Takehome exercise from last class
lfp <- read.csv('http://inta.gatech.s3.amazonaws.com/mroz_train.csv')
lfp$inlf<-as.factor(lfp$inlf) # We need the outcome variable to be a 'factor'

inlf.tree <-tree(as.numeric(inlf) - 1 ~huseduc + husage + kidslt6 + kidsge6 +
inlf.lm <- lm(as.numeric(inlf) - 1 ~huseduc + husage + kidslt6 + kidsge6 + nwi
print(inlf.tree)
# compute predictions manually, and save them as lfp$predicted.inlf
lfp$predicted.inlf<-predict(inlf.tree)


library(pROC)
evaluate <- function(y_true, y_predicted, detail=FALSE) {
  curve<-roc(y_true, y_predicted)
  if (detail) {
  print(ci.auc(curve))
  print('Sensitivity (True Positive Rate)')
  tpr<-ci.se(curve)
  print(tpr)
  print('Specificity (True Negative Rate)')
  tnr<-ci.sp(curve)
  print(tnr)
  }
  cat('Point estimate (final word on effectiveness)\n')
  print(auc(curve))
  #print('Point estimate of AUCROC: ' + auc(curve))
}
evaluate(lfp$inlf, lfp$predicted.inlf)


library(randomForest)
lfp <- read.csv('http://inta.gatech.s3.amazonaws.com/mroz_train.csv')
lfp[is.na(lfp)] <- 0
rf <-randomForest(as.factor(inlf) ~ hours  +  kidslt6  , data=lfp)
evaluate(as.factor(lfp$inlf), predict(rf,type="prob")[,1])

lfp.out <- read.csv('http://inta.gatech.s3.amazonaws.com/mroz_test.csv')
lfp.out[is.na(lfp.out)] <- 0
evaluate(as.factor(lfp.out$inlf), predict(rf, newdata=lfp.out, type="prob")[,1
evaluate(lfp.out$inlf, predict(inlf.lm, newdata=lfp.out))
evaluate(lfp.out$inlf, predict(inlf.tree, newdata=lfp.out))
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)
```

```
Call:
lm(formula = lwage ~ educ + age + married + black, data = wages)

Residuals:
     Min       1Q   Median       3Q      Max
-1.91980 -0.24114  0.01988  0.25465  1.31126

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.231197   0.159806  32.735  < 2e-16 ***
educ         0.056040   0.005820   9.629  < 2e-16 ***
age          0.019539   0.004062   4.810 1.76e-06 ***
married      0.194424   0.040952   4.748 2.38e-06 ***
black       -0.209969   0.038208  -5.495 5.03e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3834 on 930 degrees of freedom
Multiple R-squared:  0.1747,    Adjusted R-squared:  0.1711
F-statistic: 49.21 on 4 and 930 DF,  p-value: < 2.2e-16


 2SLS Estimates

Model Formula: lwage ~ educ + age + married + black

Instruments: ~feduc + age + married + black

Residuals:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-1.85005 -0.24061  0.02436  0.00000  0.26225  1.34065

               Estimate  Std. Error  t value   Pr(>|t|)
(Intercept)  4.62589626  0.26729720 17.30619 < 2.22e-16 ***
educ         0.09640575  0.01586492  6.07666 1.9681e-09 ***
age          0.02106442  0.00472040  4.46242 9.3723e-06 ***
married      0.20134756  0.04651945  4.32824 1.7109e-05 ***
black       -0.11997207  0.05366054 -2.23576   0.025667 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3944321 on 736 degrees of freedom

Installing packages into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)
```

```
Call:
lm(formula = wage ~ hours + IQ + KWW + educ + exper + tenure +
    age + married + black + south + urban + sibs, data = wages)

Residuals:
    Min      1Q  Median      3Q     Max
-849.76 -223.19  -40.18  172.09 2091.17

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -562.589    185.142  -3.039  0.00244 **
hours         -3.485      1.621  -2.150  0.03182 *
IQ             2.828      1.004   2.817  0.00495 **
KWW            5.261      1.981   2.656  0.00804 **
educ          48.213      7.182   6.713 3.33e-11 ***
exper          9.774      3.589   2.723  0.00658 **
tenure         5.242      2.406   2.178  0.02963 *
age            5.046      4.930   1.024  0.30634
married      170.231     37.883   4.494 7.89e-06 ***
black       -108.844     39.808  -2.734  0.00637 **
south        -53.689     25.536  -2.102  0.03578 *
urban        160.652     26.200   6.132 1.29e-09 ***
sibs          -1.068      5.455  -0.196  0.84488
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 352.6 on 922 degrees of freedom
Multiple R-squared:  0.2493,    Adjusted R-squared:  0.2395
F-statistic: 25.51 on 12 and 922 DF,  p-value: < 2.2e-16
```

```
Start:  AIC=10981.26
wage ~ hours + IQ + KWW + educ + exper + tenure + age + married +
    black + south + urban + sibs

          Df Sum of Sq       RSS   AIC
- sibs     1      4763 114655843 10979
- age      1    130265 114781346 10980
<none>                 114651080 10981
- south    1    549683 115200763 10984
- hours    1    574799 115225880 10984
- tenure   1    590101 115241181 10984
- KWW      1    877455 115528536 10986
- exper    1    922291 115573372 10987
- black    1    929659 115580739 10987
- IQ       1    986880 115637960 10987
- married  1   2510929 117162009 11000
- urban    1   4675218 119326298 11017
- educ     1   5603726 120254807 11024

Step:  AIC=10979.3
wage ~ hours + IQ + KWW + educ + exper + tenure + age + married +
    black + south + urban

          Df Sum of Sq       RSS   AIC
- age      1    128704 114784547 10978
<none>                 114655843 10979
- south    1    546953 115202796 10982
- hours    1    575305 115231148 10982
- tenure   1    590732 115246575 10982
- KWW      1    911607 115567450 10985
- exper    1    926789 115582632 10985
- black    1   1000335 115656178 10985
- IQ       1   1002450 115658294 10985
- married  1   2508266 117164110 10998
- urban    1   4686063 119341906 11015
- educ     1   5673927 120329770 11022

Step:  AIC=10978.35
wage ~ hours + IQ + KWW + educ + exper + tenure + married + black +
    south + urban

          Df Sum of Sq       RSS   AIC
<none>                 114784547 10978
- hours    1    562954 115347501 10981
- south    1    565737 115350283 10981
- tenure   1    680663 115465209 10982
- IQ       1    907667 115692213 10984
- black    1    980306 115764853 10984
- KWW      1   1413125 116197671 10988
- exper    1   1678662 116463208 10990
- married  1   2538972 117323519 10997
- urban    1   4639497 119424043 11013
- educ     1   6097302 120881849 11025
```

```
Call:
lm(formula = wage ~ hours + IQ + KWW + educ + exper + tenure +
    married + black + south + urban, data = wages)

Residuals:
    Min      1Q  Median      3Q     Max
-872.52 -224.26  -41.71  171.08 2086.14

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -453.2969   141.0390  -3.214 0.001354 **
hours         -3.4474     1.6194  -2.129 0.033536 *
IQ             2.6621     0.9848   2.703 0.006996 **
KWW            6.0918     1.8062   3.373 0.000775 ***
educ          49.4804     7.0627   7.006 4.73e-12 ***
exper         11.5519     3.1425   3.676 0.000251 ***
tenure         5.5777     2.3828   2.341 0.019455 *
married      171.1045    37.8476   4.521 6.96e-06 ***
black       -109.2993    38.9083  -2.809 0.005072 **
south        -54.4073    25.4951  -2.134 0.033103 *
urban        159.8942    26.1639   6.111 1.46e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 352.5 on 924 degrees of freedom
Multiple R-squared:  0.2484,    Adjusted R-squared:  0.2402
F-statistic: 30.53 on 10 and 924 DF,  p-value: < 2.2e-16
Regression tree:
tree(formula = wage ~ married + hours + IQ + KWW + educ + tenure +
    exper, data = wages, method = "anova")
Variables actually used in tree construction:
[1] "KWW"  "IQ"   "educ"
Number of terminal nodes:  6
Residual mean deviance:  130000 = 120800000 / 929
Distribution of residuals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-891.70 -250.10  -27.35    0.00  188.60 2137.00
$size
[1] 6 5 4 3 2 1

$dev
[1] 132389865 132323636 131781084 134862926 140056186 152875630

$k
[1]     -Inf  2430505  2880143  3466767  7749956 15376573

$method
[1] "deviance"

attr(,"class")
[1] "prune"         "tree.sequence"
Regression tree:
snip.tree(tree = wage.tree, nodes = 14L)
Variables actually used in tree construction:
[1] "KWW"  "IQ"   "educ"
Number of terminal nodes:  5
Residual mean deviance:  132500 = 123200000 / 930
Distribution of residuals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-891.70 -250.80  -26.53    0.00  190.40 2137.00
```

A anova: 8 × 5

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| | <int> | <dbl> | <dbl> | <dbl> | <dbl> |
| **married** | 1 | 2848893.49 | 2848893.49 | 21.7646920 | 3.535444e-06 |
| **hours** | 1 | 29758.56 | 29758.56 | 0.2273465 | 6.336108e-01 |
| **IQ** | 1 | 14964264.41 | 14964264.41 | 114.3224926 | 3.054865e-25 |
| **KWW** | 1 | 6619302.92 | 6619302.92 | 50.5694893 | 2.297498e-12 |
| **educ** | 1 | 4117977.24 | 4117977.24 | 31.4601112 | 2.686906e-08 |
| **tenure** | 1 | 1277562.58 | 1277562.58 | 9.7601950 | 1.838733e-03 |
| **exper** | 1 | 1518567.69 | 1518567.69 | 11.6014018 | 6.873314e-04 |
| **Residuals** | 927 | 121339841.32 | 130895.19 | NA | NA |

```
Regression tree:
tree(formula = wage ~ married + hours + IQ + KWW + educ + tenure +
    exper, data = wages, method = "anova")
Variables actually used in tree construction:
[1] "KWW"  "IQ"   "educ"
Number of terminal nodes:  6
Residual mean deviance:  130000 = 120800000 / 929
Distribution of residuals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-891.70 -250.10  -27.35    0.00  188.60 2137.00
```

**1:** 1468.69565217391

**1:** 1089.29299209625

```
node), split, n, deviance, yval
      * denotes terminal node

 1) root 602 148.2000 0.56150
   2) kidslt6 < 0.5 478 113.8000 0.60880
     4) husage < 48.5 273  55.2800 0.71790
       8) nwifeinc < 27.936 224  40.4600 0.76340
        16) educ < 9.5 19   4.7370 0.47370 *
        17) educ > 9.5 205  33.9800 0.79020
          34) kidsge6 < 2.5 153  20.2400 0.84310 *
          35) kidsge6 > 2.5 52  12.0600 0.63460 *
       9) nwifeinc > 27.936 49  12.2400 0.51020
        18) huseduc < 12.5 11   0.9091 0.09091 *
        19) huseduc > 12.5 38   8.8420 0.63160
          38) nwifeinc < 31.85 11   2.1820 0.27270 *
          39) nwifeinc > 31.85 27   4.6670 0.77780 *
     5) husage > 48.5 205  50.9800 0.46340
      10) educ < 12.5 155  37.2000 0.40000 *
      11) educ > 12.5 50  11.2200 0.66000 *
   3) kidslt6 > 0.5 124  29.1900 0.37900 *
Setting levels: control = 0, case = 1

Setting direction: controls < cases

Point estimate (final word on effectiveness)
Area under the curve: 0.7283
Setting levels: control = 0, case = 1

Setting direction: controls > cases

Point estimate (final word on effectiveness)
Area under the curve: 1
Setting levels: control = 0, case = 1

Setting direction: controls > cases

Point estimate (final word on effectiveness)
Area under the curve: 1
```

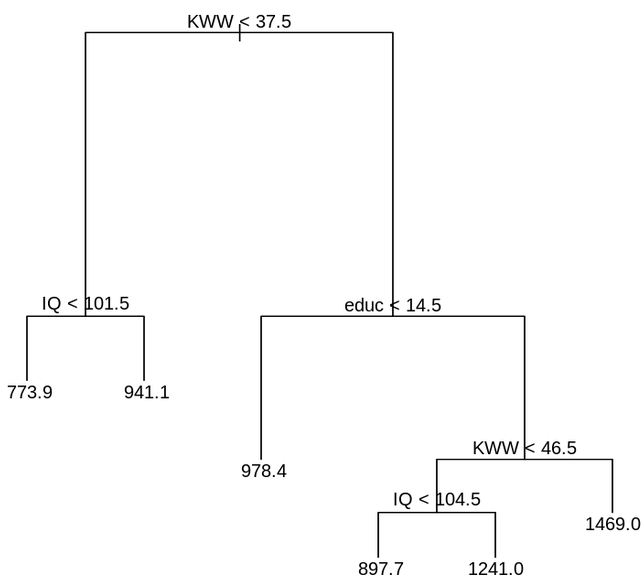
```
Setting levels: control = 0, case = 1

Setting direction: controls < cases

Point estimate (final word on effectiveness)
Area under the curve: 0.7322
Setting levels: control = 0, case = 1

Setting direction: controls < cases

Point estimate (final word on effectiveness)
Area under the curve: 0.6488
```




# Code Change

# Code Change Description

I added a train/test evaluation for the wage models. After fitting the original tree and linear model, I split the data 80/20 using set.seed(123) to keep results consistent. I refit both models on the training set and predicted wages on the held-out test set. Then I computed mean squared error (MSE) for each model and printed both numbers, plus a short note showing which model had lower test error. The original code compared models using in sample fit, but testing on held out data shows how well each model will perform on new observations. I used the same formula for both models (wage ~ married + hours + IQ + KWW + educ + tenure + exper) to keep things consistent. The change is located right after the original predict commands that compare tree and linear model predictions for a specific individual.

In [ …

```
# Instrumental Variables example
# install.packages("sem")
# Basic package setup so the script runs without manual installs
options(repos=c(CRAN='https://cloud.r-project.org'))
ensure_installed <- function(pkgs) {
  for (p in pkgs) {
    if (!requireNamespace(p, quietly=TRUE)) {
      try(install.packages(p, quiet=TRUE), silent=TRUE)
    }
  }
}
ensure_installed(c('sem','MASS','tree','pROC','randomForest','AER'))

# Try to use sem; if it isn't available, fall back to AER::ivreg
use.sem <- requireNamespace('sem', quietly=TRUE)
if (use.sem) {
  library(sem)
} else {
  library(AER)
}
wages<-read.csv('http://inta.gatech.s3.amazonaws.com/wage2.csv')
if (use.sem) {
  iv.results<-tsls(lwage ~ educ + age + married + black, ~ feduc + age + marri
} else {
  iv.results<-ivreg(lwage ~ educ + age + married + black | feduc + age + marri
}
# Run an instrumental variables regression. Note that father's education is
# used for an instrument, and is not in the first set of variables. Age,
# marital, status, and race are assumed to be less related to underlying
# ability and so are controlled for in the "first stage" regression too.

ols.results<-lm(lwage ~ educ + age + married + black, data=wages)
# Run the analogous OLS regression without father's education

print(summary(ols.results))
print(summary(iv.results))
# Note that the point estimate for education has now gone up. This suggests
# that people who have higher education in a way that's correlated with
# their father having higher education, even holding fixed age, marital
# status, and race, receive ~9% higher wages for every year of education
# Also note that the standard error on education is higher for the iv than
# ols results. The reason for this is that in IV, we're using only some of
# the variation in education: only the part related to father's education,
# so it's like there's less variation overall

# Stepwise regression and tree example
# Packages ensured above; leaving install here commented to avoid prompts
# install.packages(c('MASS','sem'))
library(MASS)
start.model<-lm(wage ~ hours + IQ + KWW + educ + exper + tenure + age + marrie
# Give an initial model, which will be the most coefficients we'd want to ever
summary(start.model)
stepwise.model<- step(start.model)
# The command "step" adds and subtracts coefficients to maximize a measure of
# goodness of fit, by default AIC
summary(stepwise.model)

# install.packages('tree')
if (!requireNamespace('tree', quietly=TRUE)) {
  try(install.packages('tree', quiet=TRUE), silent=TRUE)
}
if (!requireNamespace('tree', quietly=TRUE)) {
  stop('Package "tree" is not available. Run install.packages("tree") once, th
}
library(tree)
wage.tree <- tree(wage ~ married + hours + IQ + KWW + educ + tenure + exper, m
# fit a tree
summary(wage.tree)
# Print the results

plot(wage.tree)
text(wage.tree)
```

```
# Plot the tree, and add the text decisions

cross.validation <- cv.tree(wage.tree)
# perform cross-validation, 10-fold.

cross.validation
# look for the size with the lowest "dev" or deviance
# Why might this be different than the fitted tree?

pruned.wage.tree<-prune.tree(wage.tree,best=5)
# Prune this tree down to 5 terminal nodes, just to show this is how you would
# do it. Here this makes it worse though
summary(pruned.wage.tree)

lmfit<- lm(wage ~ married + hours + IQ + KWW + educ + tenure + exper, data=wag
# Compute a regression using those same variables:
anova(lmfit)
summary(wage.tree)
# Notice that the mean sum of squared residuals was 130895 using the linear
# model, compared to the residual mean deviance of 130000 using trees. So the
# tree method has less residual error, which means it's a better predictor.
# This is especially striking given that the tree only uses KWW, educ, IQ,
# While the regression needs substantial contributions from tenure, experience
# and marital status too.

predict(wage.tree, newdata=data.frame(IQ=100, KWW=50, educ=16, married=1, hour
predict(lmfit, newdata=data.frame(IQ=100, KWW=50, educ=16, married=1, hours=40
# These predictors give different estimates for the expected wage of a
# particular individual, relatively young, married, well educated, with good
# knowledge of the world of work. Which would you trust and why?

set.seed(123)
n <- nrow(wages)
train.index <- sample(1:n, size = floor(0.8 * n))
wages.train <- wages[train.index, ]
wages.test <- wages[-train.index, ]

# fit on training split
tree.train <- tree(wage ~ married + hours + IQ + KWW + educ + tenure + exper,
lm.train <- lm(wage ~ married + hours + IQ + KWW + educ + tenure + exper, data

# predict on test split and compute MSE
pred.tree <- predict(tree.train, newdata=wages.test)
pred.lm <- predict(lm.train, newdata=wages.test)
mse.tree <- mean((wages.test$wage - pred.tree)^2, na.rm=TRUE)
mse.lm <- mean((wages.test$wage - pred.lm)^2, na.rm=TRUE)

cat('Test MSE (tree):', round(mse.tree, 2), '\n')
cat('Test MSE (linear model):', round(mse.lm, 2), '\n')
if (!is.na(mse.tree) && !is.na(mse.lm)) {
  better <- ifelse(mse.tree < mse.lm, 'tree', 'linear model')
  cat('Lower test error from:', better, '\n')
}

# Extension: try to build another tree, and see what it looks like.
# You can sample the data with the command:
# install.packages('dplyr')
# library('dplyr')
# wages.sample <- sample(wages, n=500)

# Takehome exercise from last class
lfp <- read.csv('http://inta.gatech.s3.amazonaws.com/mroz_train.csv')
lfp$inlf<-as.factor(lfp$inlf) # We need the outcome variable to be a 'factor'

inlf.tree <-tree(as.numeric(inlf) - 1 ~huseduc + husage + kidslt6 + kidsge6 +
inlf.lm <- lm(as.numeric(inlf) - 1 ~huseduc + husage + kidslt6 + kidsge6 + nwi
print(inlf.tree)
# compute predictions manually, and save them as lfp$predicted.inlf
lfp$predicted.inlf<-predict(inlf.tree)


library(pROC)
evaluate <- function(y_true, y_predicted, detail=FALSE) {
```

```
  curve<-roc(y_true, y_predicted)
  if (detail) {
  print(ci.auc(curve))
  print('Sensitivity (True Positive Rate)')
  tpr<-ci.se(curve)
  print(tpr)
  print('Specificity (True Negative Rate)')
  tnr<-ci.sp(curve)
  print(tnr)
  }
  cat('Point estimate (final word on effectiveness)\n')
  print(auc(curve))
  #print('Point estimate of AUCROC: ' + auc(curve))
}
evaluate(lfp$inlf, lfp$predicted.inlf)


library(randomForest)
lfp <- read.csv('http://inta.gatech.s3.amazonaws.com/mroz_train.csv')
lfp[is.na(lfp)] <- 0
rf <-randomForest(as.factor(inlf) ~ hours  +  kidslt6  , data=lfp)
evaluate(as.factor(lfp$inlf), predict(rf,type="prob")[,1])

lfp.out <- read.csv('http://inta.gatech.s3.amazonaws.com/mroz_test.csv')
lfp.out[is.na(lfp.out)] <- 0
evaluate(as.factor(lfp.out$inlf), predict(rf, newdata=lfp.out, type="prob")[,1
evaluate(lfp.out$inlf, predict(inlf.lm, newdata=lfp.out))
evaluate(lfp.out$inlf, predict(inlf.tree, newdata=lfp.out))
```

```
also installing the dependencies ‘cowplot’, ‘Deriv’, ‘microbenchmark’, ‘numDe
riv’, ‘doBy’, ‘SparseM’, ‘MatrixModels’, ‘carData’, ‘pbkrtest’, ‘quantreg’,
‘car’, ‘lmtest’, ‘sandwich’, ‘zoo’, ‘Formula’
```

```
Call:
lm(formula = lwage ~ educ + age + married + black, data = wages)

Residuals:
     Min       1Q   Median       3Q      Max
-1.91980 -0.24114  0.01988  0.25465  1.31126

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.231197   0.159806  32.735  < 2e-16 ***
educ         0.056040   0.005820   9.629  < 2e-16 ***
age          0.019539   0.004062   4.810 1.76e-06 ***
married      0.194424   0.040952   4.748 2.38e-06 ***
black       -0.209969   0.038208  -5.495 5.03e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3834 on 930 degrees of freedom
Multiple R-squared:  0.1747,    Adjusted R-squared:  0.1711
F-statistic: 49.21 on 4 and 930 DF,  p-value: < 2.2e-16


 2SLS Estimates

Model Formula: lwage ~ educ + age + married + black

Instruments: ~feduc + age + married + black

Residuals:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-1.85005 -0.24061  0.02436  0.00000  0.26225  1.34065

               Estimate  Std. Error  t value   Pr(>|t|)
(Intercept)  4.62589626  0.26729720 17.30619 < 2.22e-16 ***
educ         0.09640575  0.01586492  6.07666 1.9681e-09 ***
age          0.02106442  0.00472040  4.46242 9.3723e-06 ***
married      0.20134756  0.04651945  4.32824 1.7109e-05 ***
black       -0.11997207  0.05366054 -2.23576   0.025667 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3944321 on 736 degrees of freedom

Call:
lm(formula = wage ~ hours + IQ + KWW + educ + exper + tenure +
    age + married + black + south + urban + sibs, data = wages)

Residuals:
    Min      1Q  Median      3Q     Max
-849.76 -223.19  -40.18  172.09 2091.17

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -562.589    185.142  -3.039  0.00244 **
hours         -3.485      1.621  -2.150  0.03182 *
IQ             2.828      1.004   2.817  0.00495 **
KWW            5.261      1.981   2.656  0.00804 **
educ          48.213      7.182   6.713 3.33e-11 ***
exper          9.774      3.589   2.723  0.00658 **
tenure         5.242      2.406   2.178  0.02963 *
age            5.046      4.930   1.024  0.30634
married      170.231     37.883   4.494 7.89e-06 ***
black       -108.844     39.808  -2.734  0.00637 **
south        -53.689     25.536  -2.102  0.03578 *
urban        160.652     26.200   6.132 1.29e-09 ***
sibs          -1.068      5.455  -0.196  0.84488
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 352.6 on 922 degrees of freedom
Multiple R-squared:  0.2493,    Adjusted R-squared:  0.2395
F-statistic: 25.51 on 12 and 922 DF,  p-value: < 2.2e-16
```

```
Start:  AIC=10981.26
wage ~ hours + IQ + KWW + educ + exper + tenure + age + married +
    black + south + urban + sibs

          Df Sum of Sq       RSS   AIC
- sibs     1      4763 114655843 10979
- age      1    130265 114781346 10980
<none>                 114651080 10981
- south    1    549683 115200763 10984
- hours    1    574799 115225880 10984
- tenure   1    590101 115241181 10984
- KWW      1    877455 115528536 10986
- exper    1    922291 115573372 10987
- black    1    929659 115580739 10987
- IQ       1    986880 115637960 10987
- married  1   2510929 117162009 11000
- urban    1   4675218 119326298 11017
- educ     1   5603726 120254807 11024

Step:  AIC=10979.3
wage ~ hours + IQ + KWW + educ + exper + tenure + age + married +
    black + south + urban

          Df Sum of Sq       RSS   AIC
- age      1    128704 114784547 10978
<none>                 114655843 10979
- south    1    546953 115202796 10982
- hours    1    575305 115231148 10982
- tenure   1    590732 115246575 10982
- KWW      1    911607 115567450 10985
- exper    1    926789 115582632 10985
- black    1   1000335 115656178 10985
- IQ       1   1002450 115658294 10985
- married  1   2508266 117164110 10998
- urban    1   4686063 119341906 11015
- educ     1   5673927 120329770 11022

Step:  AIC=10978.35
wage ~ hours + IQ + KWW + educ + exper + tenure + married + black +
    south + urban

          Df Sum of Sq       RSS   AIC
<none>                 114784547 10978
- hours    1    562954 115347501 10981
- south    1    565737 115350283 10981
- tenure   1    680663 115465209 10982
- IQ       1    907667 115692213 10984
- black    1    980306 115764853 10984
- KWW      1   1413125 116197671 10988
- exper    1   1678662 116463208 10990
- married  1   2538972 117323519 10997
- urban    1   4639497 119424043 11013
- educ     1   6097302 120881849 11025
```

```
Call:
lm(formula = wage ~ hours + IQ + KWW + educ + exper + tenure +
    married + black + south + urban, data = wages)

Residuals:
    Min      1Q  Median      3Q     Max
-872.52 -224.26  -41.71  171.08 2086.14

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -453.2969   141.0390  -3.214 0.001354 **
hours         -3.4474     1.6194  -2.129 0.033536 *
IQ             2.6621     0.9848   2.703 0.006996 **
KWW            6.0918     1.8062   3.373 0.000775 ***
educ          49.4804     7.0627   7.006 4.73e-12 ***
exper         11.5519     3.1425   3.676 0.000251 ***
tenure         5.5777     2.3828   2.341 0.019455 *
married      171.1045    37.8476   4.521 6.96e-06 ***
black       -109.2993    38.9083  -2.809 0.005072 **
south        -54.4073    25.4951  -2.134 0.033103 *
urban        159.8942    26.1639   6.111 1.46e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 352.5 on 924 degrees of freedom
Multiple R-squared:  0.2484,    Adjusted R-squared:  0.2402
F-statistic: 30.53 on 10 and 924 DF,  p-value: < 2.2e-16
Regression tree:
tree(formula = wage ~ married + hours + IQ + KWW + educ + tenure +
    exper, data = wages, method = "anova")
Variables actually used in tree construction:
[1] "KWW"  "IQ"   "educ"
Number of terminal nodes:  6
Residual mean deviance:  130000 = 120800000 / 929
Distribution of residuals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-891.70 -250.10  -27.35    0.00  188.60 2137.00
$size
[1] 6 5 4 3 2 1

$dev
[1] 128841719 132123890 132407768 132002303 138398206 153248256

$k
[1]     -Inf  2430505  2880143  3466767  7749956 15376573

$method
[1] "deviance"

attr(,"class")
[1] "prune"         "tree.sequence"
Regression tree:
snip.tree(tree = wage.tree, nodes = 14L)
Variables actually used in tree construction:
[1] "KWW"  "IQ"   "educ"
Number of terminal nodes:  5
Residual mean deviance:  132500 = 123200000 / 930
Distribution of residuals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-891.70 -250.80  -26.53    0.00  190.40 2137.00
```

A anova: 8 × 5

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| | <int> | <dbl> | <dbl> | <dbl> | <dbl> |
| **married** | 1 | 2848893.49 | 2848893.49 | 21.7646920 | 3.535444e-06 |
| **hours** | 1 | 29758.56 | 29758.56 | 0.2273465 | 6.336108e-01 |
| **IQ** | 1 | 14964264.41 | 14964264.41 | 114.3224926 | 3.054865e-25 |
| **KWW** | 1 | 6619302.92 | 6619302.92 | 50.5694893 | 2.297498e-12 |
| **educ** | 1 | 4117977.24 | 4117977.24 | 31.4601112 | 2.686906e-08 |
| **tenure** | 1 | 1277562.58 | 1277562.58 | 9.7601950 | 1.838733e-03 |
| **exper** | 1 | 1518567.69 | 1518567.69 | 11.6014018 | 6.873314e-04 |
| **Residuals** | 927 | 121339841.32 | 130895.19 | NA | NA |

```
Regression tree:
tree(formula = wage ~ married + hours + IQ + KWW + educ + tenure +
    exper, data = wages, method = "anova")
Variables actually used in tree construction:
[1] "KWW"  "IQ"   "educ"
Number of terminal nodes:  6
Residual mean deviance:  130000 = 120800000 / 929
Distribution of residuals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-891.70 -250.10  -27.35    0.00  188.60 2137.00
```

**1:** 1468.69565217391

**1:** 1089.29299209625

```
Test MSE (tree): 187127.5
Test MSE (linear model): 170416.1
Lower test error from: linear model
node), split, n, deviance, yval
      * denotes terminal node

 1) root 602 148.2000 0.56150
   2) kidslt6 < 0.5 478 113.8000 0.60880
     4) husage < 48.5 273  55.2800 0.71790
       8) nwifeinc < 27.936 224  40.4600 0.76340
        16) educ < 9.5 19   4.7370 0.47370 *
        17) educ > 9.5 205  33.9800 0.79020
          34) kidsge6 < 2.5 153  20.2400 0.84310 *
          35) kidsge6 > 2.5 52  12.0600 0.63460 *
       9) nwifeinc > 27.936 49  12.2400 0.51020
        18) huseduc < 12.5 11   0.9091 0.09091 *
        19) huseduc > 12.5 38   8.8420 0.63160
          38) nwifeinc < 31.85 11   2.1820 0.27270 *
          39) nwifeinc > 31.85 27   4.6670 0.77780 *
     5) husage > 48.5 205  50.9800 0.46340
      10) educ < 12.5 155  37.2000 0.40000 *
      11) educ > 12.5 50  11.2200 0.66000 *
   3) kidslt6 > 0.5 124  29.1900 0.37900 *
Type 'citation("pROC")' for a citation.


Attaching package: ‘pROC’


The following objects are masked from ‘package:stats’:

    cov, smooth, var


Setting levels: control = 0, case = 1

Setting direction: controls < cases

Point estimate (final word on effectiveness)
Area under the curve: 0.7283
```

```
randomForest 4.7-1.2

Type rfNews() to see new features/changes/bug fixes.

Setting levels: control = 0, case = 1

Setting direction: controls > cases

Point estimate (final word on effectiveness)
Area under the curve: 0.9999
Setting levels: control = 0, case = 1

Setting direction: controls > cases

Point estimate (final word on effectiveness)
Area under the curve: 1
Setting levels: control = 0, case = 1

Setting direction: controls < cases

Point estimate (final word on effectiveness)
Area under the curve: 0.7322
Setting levels: control = 0, case = 1

Setting direction: controls < cases

Point estimate (final word on effectiveness)
Area under the curve: 0.6488
```